SNP imputation in association studies

Eran Halperin & Dietrich A Stephan

Only a subset of single-nucleotide polymorphisms (SNPs) can be genotyped in genome-wide association studies. Imputation methods can infer the alleles of 'hidden' variants and use those inferences to test the hidden variants for association.

The large amount of data generated in whole-genome association studies, involving hundreds of thousands of SNPs genotyped in thousands of individuals, complicates the statistical and computational analysis of that data. The correlation between SNPs (linkage disequilibrium) enables much of the variation to be captured despite the inability to genotype all SNPs, and our previous primer[1] described how tagSNPs and haplotypes have been used as proxies for neighboring associations. However, especially with the advent of high-throughput genotyping technologies, the key challenge has started to shift from identifying tagSNPs that best capture genetic variation in the population to the ability to interrogate SNPs not covered by these technologies. Moreover, how does one consolidate distinct data sets when subsets of the same population are genotyped with slightly different technologies that have different capacities?

Imputation methods address these problems by using the linkage disequilibrium structure in a region to infer the alleles of SNPs not directly genotyped in the study (hidden SNPs). The starting point of imputation methods is a reference data set such as the HapMap, in which a large set of SNPs is being genotyped. The underlying assumption is that the reference samples, the cases and the controls are all sampled from the same population. Under this assumption, the three populations share the same linkage disequilibrium structure and the same haplotype distribution for every set of SNPs. Thus, the structure of the linkage disequilibrium in the reference population, in conjunction with the structure of the linkage disequilibrium of the observed SNPs within both the cases and the controls, can be used to impute the alleles of a hidden SNP. Imputed SNPs can then be tested for association using an appropriate statistical test.

The rationale that underlies imputation methods is that even though the causal SNP may not have been genotyped in the study at hand, it may have been genotyped in the reference population. In this case, simulations have revealed that the imputation of SNPs that appear in the reference population facilitates detection of association. Imputation methods are also invaluable when multiple data sets,

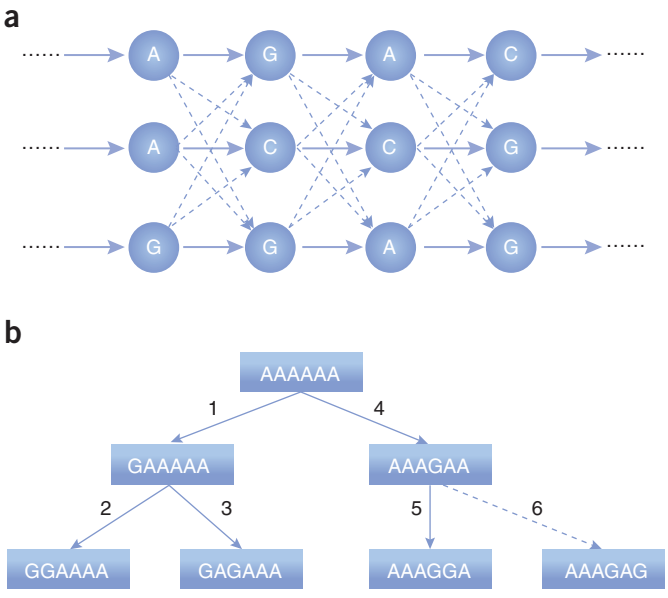

Figure 1 Models used in imputation methods. (**a**) A generic Hidden Markov Model used for imputation and phasing. Every circle is a state, each column corresponds to a SNP and each row corresponds to an ancestral haplotype. According to this model, a haplotype is generated by a random walk on the Markov chain from left to right, where the transition probabilities from one haplotype to another (denoted by the dashed arrows) are determined by the recombination rate and physical distance between the two SNPs. At each position, there is a small probability that the resulting haplotype will be mutated further. A genotype is generated at the conjunction of two such haplotypes. (**b**) A perfect phylogeny tree explaining the genealogy of the haplotypes, and leading to a test of the hidden SNP 6. Each node in the tree corresponds to a haplotype, and each edge corresponds to a mutating position. A perfect phylogeny model assumes no recurrent mutations or recombination events. The dashed line corresponds to an unobserved SNP (at position 6), which can be tested for association by testing the haplotypes spanned by SNPs 4 and 5.

Eran Halperin is at the International Computer Science Institute, Berkeley, California, USA, and the Blavatnik School of Computer Science and the Department of Biotechnology, Tel Aviv University, Tel Aviv, Israel. Dietrich A. Stephan is in the Division of Genomics Research, Navigenics, Foster City, California, USA. e-mail: dietrich@navigenics.com

Table 1 Comparisons of imputation methods

Software	Model	Uses reference?	Optimization method
IMPUTE[3]	Hidden Markov Model	Yes	Markov Chain Monte Carlo
MACH[7]	Hidden Markov Model	Yes	Iteratively assigns haplotypes to the genotypes based on the converging model
BIMBAM[13] (FastPHASE[5])	Hidden Markov Model	Yes	Uses a small number of states (haplotype clustering)
TUNA[2]	Weighted haplotype proxies	Yes	Greedily searching for the proxies
SNPMStat[6]	Likelihood-based diplotype proxies	Yes	Maximizes likelihood based on the possible diplotypes explaining the genotype
CAMP[10]	Coalescent	No	Builds an approximate perfect-phylogeny tree

obtained using different genotyping platforms that span different sets of SNPs, need to be consolidated into a single meta-analysis. As some SNPs are present in one data set and not in the others, a naive approach can at best only hope to gain power for those SNPs genotyped in more than one study. Imputation can increase power for the SNPs participating in the union of the studies.

Methods for haplotype proxies follow a similar rationale, as described in a companion primer[1]. Imputation methods, such as TUNA[2], are based on haplotype proxies, and more direct approaches for imputations, based on variants of Hidden Markov Models (HMM) in which the haplotype structure is used implicitly, have been suggested more recently[3–7]. The basic model comprises a few hidden states per SNP, representing the possible ancestral haplotypes at the SNP. The assumption is that the alleles across a chromosome (haplotypes) are generated by a random walk across these states, where the transition probability from one state to another depends on the population-scaled recombination rate (ρ) between the two SNPs (**Fig. 1**). Each state corresponds to an allele, however, with some small probability the haplotype's allele could differ from the state's allele, representing a mutation or a genotyping error. The different imputation methods differ mainly in the way in which the ancestral haplotypes are chosen, in the assignments of transition probabilities and in the optimization procedures used.

The HMM-based imputation methods can be distinguished by the way the Markov chain's states are defined, as well as by how the parameters of the Markov chain are learned (**Table 1**). Whereas Fastphase[5] defines a Markov model with a small number of states at each position, IMPUTE[3] and MACH[7] each have a very large state set. The parameters of the Markov chain can be estimated using integration; if the haplotypes were known, we could find the transition probabilities of the chain by a simple counting procedure. Even if they are not known, we can estimate the parameters by taking the average across all possible pairs of haplotypes that are compatible with the genotype. Methods such as IMPUTE and Fastphase use Markov Chain Monte Carlo methods to perform this integration. In contrast, MACH performs a local search, leveraged on the long identical stretches of DNA still shared by unrelated individuals.

Many of the imputation methods are based on ideas that were previously developed in the context of phasing, or haplotype inference from genotypes. For example, Fastphase[5] and Gerbil[8] are phasing methods based on HMMs like the one shown in **Figure 1a**. To date, the most accurate methods for phasing are based on genealogical models, in which the genealogical tree of the haplotypes is used to constrain dependencies. For instance, the method PHASE[9] assumes that the haplotypes are generated by a genealogical tree that follows the coalescent model; it samples haplotypes and genealogies conditioned on the genotype information.

Genealogy-based methods have been suggested in cases where imputation is implicit. If the causal SNP has not been genotyped in the reference population, or if there is no reliable reference population (that is, one derived from the exact same population of the cases and the controls), one can still use the genealogy of the haplotypes to implicitly impute the causal SNP and then test it. For instance, Kimmel *et al.*[10] suggest approximating the genealogy using a perfect phylogeny tree, which explains the order of mutations throughout history under the assumption of no recurrence mutations and no recombination events. They construct such a tree locally for each region in the genome; subsequently, the causal SNP is characterized by an unobserved branch in the tree (**Fig. 1b**), which can be translated into a statistical test involving multiple SNPs in the region. Other methods, such as that of Minichiello and Durbin[11], aim at a more realistic model that includes the modeling of recombination events. However, Kimmel *et al.*[10] show that this approach in fact reduces power, probably owing to inaccuracies in the genealogy reconstruction and an increased multiple hypotheses burden.

An advantage of genealogy-based methods is that they do not rely on the assumption that there exists a reference data set which is sampled from the exact same population as the studied population. In fact, the choice of the reference data set has a tremendous effect on the accuracy of the imputation[12]. First, if the reference data set is too small, imputation methods will over-fit the data, which in turn affects the accuracy of the imputation. In this regard, it seems that the limited number of individuals (30 trios of European descent) genotyped in the current HapMap phase II data set may limit improvements of imputation accuracy. Second, Huang *et al.*[12] show that the fraction of missing data in the studied data set has a considerable effect on the imputation accuracy, and it can reduce the accuracy from 90–95% to 80–85%[12] when half of the data is missing. Finally, they show that the population of the reference data set plays an important role in the accuracy of the imputation. The accuracy of the imputation drops considerably, especially when the reference population and the study population differ. Moreover, it turns out that when trying to impute genotype information on some populations for which a reference population does not exist (e.g., populations from central Asia or the Middle East), it is beneficial to use a mixed reference population with different proportions of the HapMap data sets (that is, a mix of European, African and Asian populations).

The possibility of imputation errors must be considered when testing an imputed SNP for association[4]. Particular care should be taken when analyzing SNPs with low minor allele frequency, as imputation methods tend to be less accurate for such SNPs. Ultimately, the null hypothesis is invalid in any test that does not consider the accuracy of the imputation, and therefore, such analyses may result in an excess of false positives and loss of power. To account for this, a few methods (e.g. refs. 3,7,13) incorporate the uncertainty in the imputation result into the association test. Servin and Stephens[13] propose a Bayesian framework for association, which takes into account the uncertainty in imputation by averaging across the distribution of the imputed alleles. Although their framework is more robust than other frameworks that do not take into account the uncertainty in imputation, it is still advisable to genotype the associated imputed SNPs as a follow-up.

A particularly successful implementation of imputation identified 21 new associations for Crohn's disease following meta-analysis of three data sets[14]. The imputed SNPs were then genotyped across the three populations to ensure that the associations were not due to imputation errors. As only 11 associations were known before this study, this exemplifies the potential of imputation to enhance the power of association studies and underscores the importance of rigorous and efficient methods for the analysis and interpretation of association studies.

ACKNOWLEDGMENTS

E.H. is a faculty fellow of the Edmond J. Safra Bioinformatics program at Tel Aviv University. E.H. was supported by National Science Foundation grant IIS-071325412 and by the Israel Science Foundation grant no. 04514831.

1. Halperin, E. & Stephan, D.A. *Nat. Biotechnol.* **27**, 255–256 (2009).
2. Nicolae, D.L. *Genet. Epidemiol.* **30**, 718–727 (2006).
3. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. *Nat. Genet.* **39**, 906–913 (2007).
4. Pei, Y.F., Li, J., Zhang, L., Papasian, C.J. & Deng, H.W. *PLoS One* **3**, e3551 (2008).
5. Scheet, P. & Stephens, M. *Am. J. Hum. Genet.* **78**, 629–644 (2006).
6. Lin, D.Y., Hu, Y. & Huang, B.E. *Am. J. Hum. Genet.* **82**, 444–452 (2008).
7. Li, Y. & Abecasis, G.R. *Am. J. Hum. Genet.* **S79**, 2290 (2006).
8. Kimmel, G. & Shamir, R. Proc. *Natl. Acad. Sci.* USA **102**, 158–162 (2005).
9. Stephens, M., Smith, N.J. & Donnelly, P. *Am. J. Hum. Genet.* **68**, 978–989 (2001).
10. Kimmel, G., Karp, R.M., Jordan, M.I. & Halperin, E. *Am. J. Hum. Genet.* **83**, 675–683 (2008).
11. Minichiello, M.J. & Durbin, R. *Am. J. Hum. Genet.* **79**, 910–922 (2006).
12. Huang, L. *et al. Am. J. Hum. Genet.* **84**, 235–250 (2009).
13. Servin, B. & Stephens, M. *PLoS Genet.* **3**, e114 (2007).
14. Barrett, J.C. *et al. Nat. Genet.* **40**, 955–962 (2008).